MASSACHUSETTS INSTITUTE OF TECHNOLOGY

ARTIFICIAL INTELLIGENCE LABORATORY

Memo No. 324

December 1974

ON THE PURPOSE OF LOW-LEVEL VISION

by

David Marr

ABSTRACT

This article advances the thesis that the purpose of low-level vision is to encode symbolically all of the useful information contained in an intensity array, using a vocabulary of very low-level symbols: subsequent processes should have access only to this symbolic description. The reason is one of computational expediency: it allows the low-level processes to run almost autonomously; and it greatly simplifies the application of criteria to an image, whose representation in terms of conditions on the initial intensities, or on simple measurements made from them, is very cumbersome. The implications of this thesis for physiological and for computational approaches to vision are discussed. A list is given of several computational problems in low-level vision: some of these are dealt with in the accompanying articles.

Work reported herein was conducted at the Artificial Intelligence Laboratory, a Massachusetts Institute of Technology research program supported in part by the Advanced Research Projects Agency of the Department of Defense and monitored by the Office of Naval Research under Contract number N00014-70-A-0362-0005.

Introduction

Our ready appreciation of "sketches" indicates that we find a symbolic representation of the visual world particularly congenial; even a single stroke can convey a complex meaning with a startling sense of immediacy. This shows that we are capable of rapid comprehension of material only very abstractly related to the raw data on which our perceptions are based. It is difficult to believe that the artful selection and subtle manipulation of powerful visual symbols is a superficial device; the natural presumption is that it is a fundamental characteristic of the computations by which we interpret diverse kinds of sensory information, and combine them to maintain a perceptual model of the outside world.

In its extreme form, this presumption implies that what we call the "perception" of an object or state of affairs corresponds rather directly to the making, in some central place, of one or more abstract symbolic assertions about that object; and to the consequent availability of other knowledge related to that percept. The exact nature of the assertions that are made in a given circumstance will depend upon how the information is to be used, because of the fundamentally utilitarian nature of the process. Attractive though this view is as a starting point for a theory of perception, one cannot argue convincingly for or against it without having some idea whether a system that is built along these lines can work at all, and if so, how well. Until one has insight into

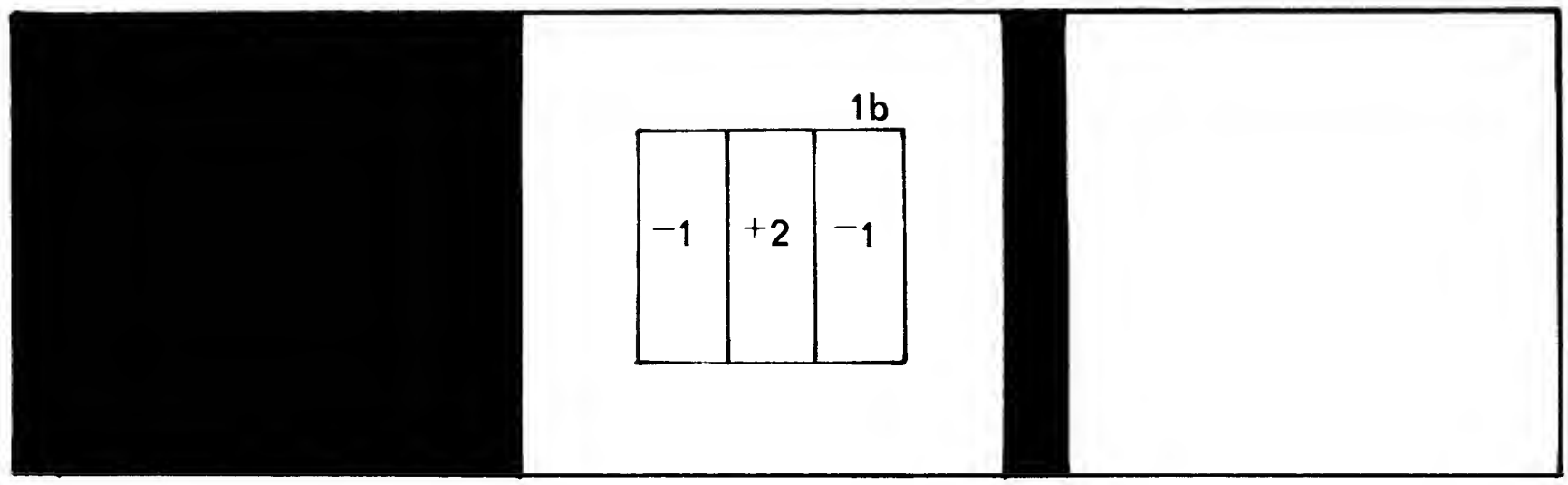
the nature of the computational problems that the nervous system so effortlessly solves, it is unlikely that we shall understand how it solves them, except in rare or in simple circumstances. The business of a theoretical investigation is therefore firstly, to explore the necessary underlying structure of the computations that need to be performed in order that high-level assertions may be made about the world; and secondly, to formulate criteria by which the implementation of such computations may be identified.

Measurements and symbols

This article introduces some of the early problems that are raised by this position in the context of the processing of visual information. It is inevitable that the analysis of sensory information should start with the making of measurements upon the incoming data; and one would like to be fairly clear about the stage at which a transformational style of computation ends, and symbolic manipulations begin. Our first task is therefore to distinguish between the concepts of a measurement, a symbolic assertion, and the representation in a computing machine of a symbolic assertion. For present purposes, a measurement will be regarded as the result of applying a function to some domain, and it is a number. When one reads a weighing machine, the act of translating the position of its pointer into a number is from this point of view an act of measurement. A symbolic assertion is not a number: it is a list or sequence of one or more atomic symbols, drawn from a vocabulary whose power derives from the conventions according to which

those symbols are used. (WEIGHT 165lbs) is an example of an assertion for which the measurement from a weighing machine can provide reliable evidence; but it is not the only possible one, and it is true only if the machine is in good order. The measurement that is actually made concerns the displacement of a pointer, and an assertion about the position of that pointer is the only thing that a disappointed dieter would be forced to accept from the evidence.

The representation of a symbolic assertion is a concept that has meaning only in the context of a computational machine that can manipulate symbolic assertions; and in that context, it may be defined as the form in which an assertion is made immediately available to a process that can use it. Because the result of any measurement may be identified with an assertion of the form (VALUE f x), where f is the name of the function that was applied and x is the name of its value, there is always at least one assertion associated with a measurement: I shall refer to this as the trivial assertion, and it is the one that a sceptic is forced to accept. The usefulness of a measurement arises because it leads either immediately, or after some computation, to a non-trivial assertion, which in the case of the weighing machine concerns the weight of whatever is on its platform. When an assertion follows immediately from a measurement, the representation of the result of the measurement may be identified with the assertion that it reliably supports. Inside a computing machine, these identifications are made when the result of a measurement is treated by subsequent processes as if it were a non-trivial assertion. Such identifications are central to the structure of a



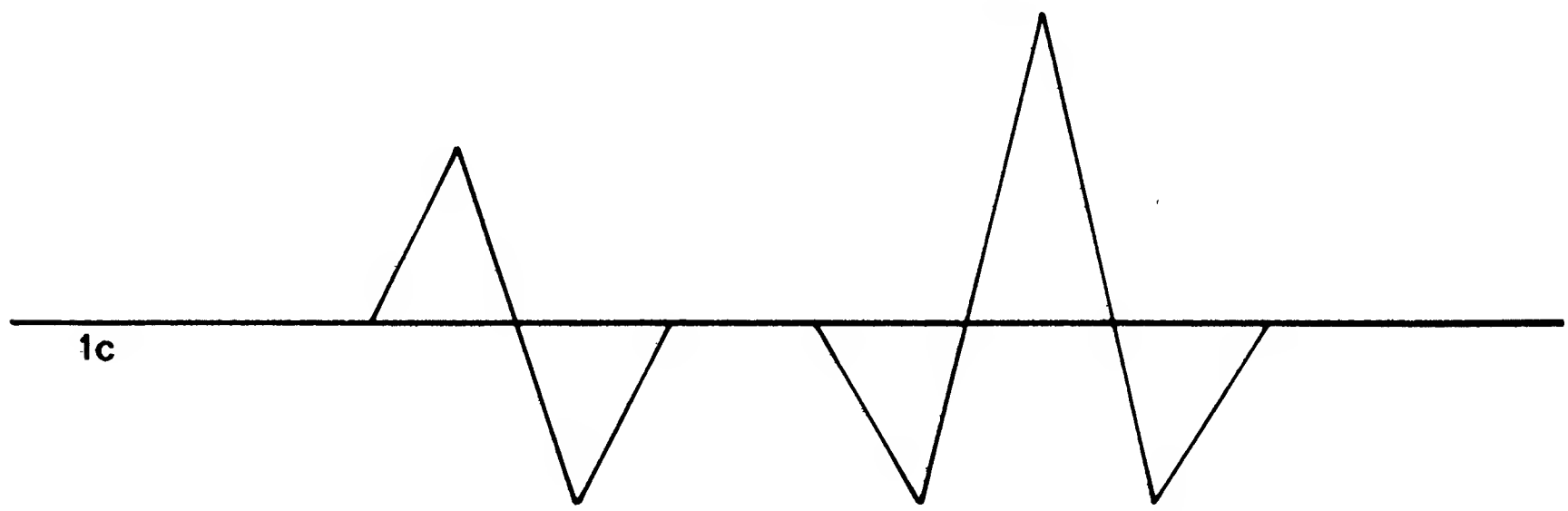


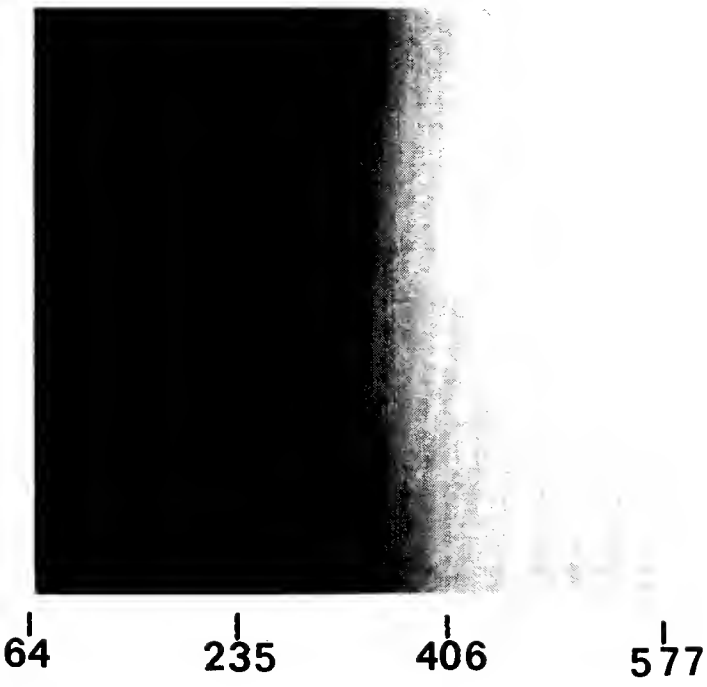
FIGURE 1

Figure 1. The image 1a shows step changes in intensity. 1c shows the convolution of the bar mask 1b with the image. This illustrates why the response of a bar-type simple cell cannot be thought of as representing an assertion about a BAR in the image.

computation; and they are virtually impossible to infer from a working program unless one has a detailed knowledge of the nature of the computation being performed by it.

The usefulness of the distinction

Simple cells in the cat make measurements upon an image, and the nature of the measurement that they make is fairly well understood. Their receptive fields are either bar- or edge-shaped (Hubel & Wiesel 1962), and if other parameters are held constant, they signal the linear convolution of a bar- or edge-shaped mask with the intensity distribution currently falling upon the retina, in logarithmic units of contrast (Maffei & Fiorentini 1973 figure 8). The important question for understanding the analysis of visual information is whether these cells represent assertions other than the trivial one associated with the measurement; and if they do, what are they? Few investigators have been incautious enough to suggest that a bar-type simple cell represents an assertion about the presence of a bar in the visual field. Figure 1, which shows the results of a bar-mask convolution with variously configured step changes in intensity, illustrates why such caution is well-founded: such a mechanism operating at an isolated edge in the image, would provoke two assertions, about a light and about a dark bar; this description is manifestly misleading. Furthermore, to produce even that misleading interpretation, numerous other problems would have to be overcome, concerning the neglect of cells at neighbouring positions and orientations to those giving the largest signals. Figure 2 shows an



2a

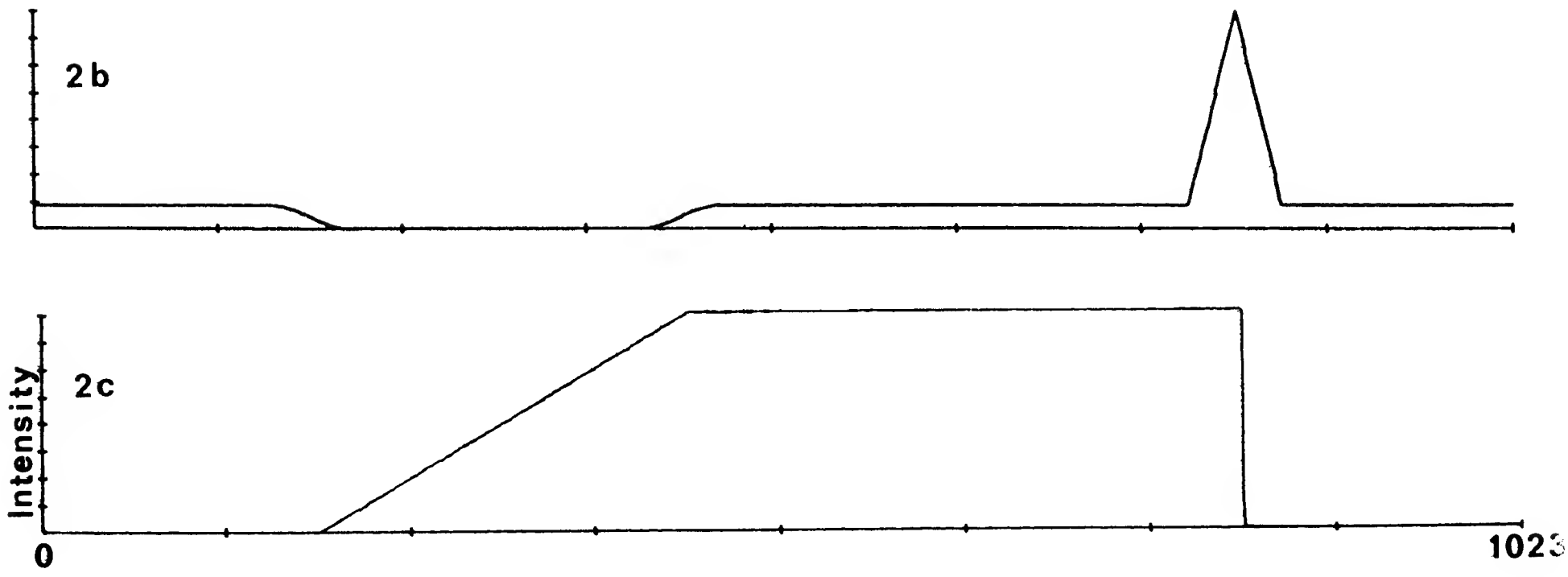


FIGURE 2

Figure 2. The image 2a has an intensity distribution the log of whose intensity is a linear function (2c) of the x-coordinate. 2b shows the convolution of this distribution with an edge mask. This illustrates why an edge-type simple cell is not immediately identifiable with an assertion about the presence of a small, sharp edge in the image.

intensity distribution whose logarithm is a linear function of the position coordinate x. A simple cell with an edge-shaped receptive field would give a constant signal along the x axis, because it measures an approximation to the local intensity gradient: yet there are no small, sharp, faint edges in the image.

In order to understand in any deep sense the function of simple cells, one needs to know two things: firstly, what does the vocabulary of assertions look like, for whose computation simple cell measurements are used; and secondly, how may these computations be characterised. There are two extreme kinds of answer to the first question. The first is that the earliest non-trivial assertions to be computed are very high level ones, using predicates like CHAIR, LION, and so forth. One could conceive of ways of doing this: one might for example use a battery of specialised filters in the Fourier domain, as has been done for the recognition of stereotyped shapes on a printed page. But this technique fails for more natural images, because the appearance of an object is an inconstant and ephemeral phenomenon.

The opposite extreme would occur if the assertions that were computed immediately were very low-level ones. I take the latter position, and will argue that it is computationally expedient to encode all of the useful information in an intensity array in symbolic form immediately. The number of different symbols that are required to cover most situations is not as large as one might have expected Marr (1974a and 1974b).

The second question is how to characterise the computations that

can lead to these low-level assertions. One can of course specify an algorithm; but although an investigation of this kind is under an obligation to describe working algorithms, its more important function is to provide an abstract characterisation of the computation that the algorithms perform. I follow Jardine & Sibson (1971) in calling such an abstract characterisation a method. The particular choice of an algorithm for implementing a method depends upon one's resources and circumstances: a neurophysiologist may for example be more interested in parallel than in serial algorithms, whereas this would not be true of a conventional programmer. The important point is that the distinction between a serial and a parallel algorithm is often a superficial matter: a characterisation in terms of methods allows us to work at a deeper level.

The distinction is an important one, because it is only at the implementation level that neurophysiological experiments can be carried out. Such experiments may be able to reject algorithms, but only rarely will they provide disproof of methods. For example, the deep issue that this article raises is whether the nervous system uses edge assertions in its computations, and by implication, whether it uses other more complex assertions on its way through the recognition process. Whether such assertions are coded for by single cells is an interesting, but lesser issue. Neurophysiological experiment may be able to test the latter, but it does not at present have the tools even to ask the former question. Computational experiments, on the other hand, can lead to a proof that a method cannot work, because they allow one to examine the reasons for its failure in a particular circumstance. Progress in understanding the

findings of visual neurophysiologists will therefore depend to a large extent upon the insights that we are able to provide into the computational problems that are involved.

Computer vision research

The most progress in machine vision has been made in the field of scene analysis, a line of development that was originated by Guzman (1968), pursued by Huffman (1970) and by Clowes (1971), and recently reached what is probably its final form in the work of Waltz (1972) and of Mackworth (1973). Scene analysis is the process of turning a line drawing of a collection of objects into a description of the objects and their relative positions. Waltz's program is a graphic example of the truth of an old adage: that the central problem of vision is how to bring the right knowledge to bear at the right time and in the right way. It incidentally emphasises the fact that the amount of knowledge that a working vision system will need is very large. Waltz showed that, when the line drawings are complete, and when the database of knowledge about what is and what is not possible in a line drawing is also complete, the sum of the constraints imposed often forces a unique interpretation on the drawing.

Admirable though this demonstration is, it is not generally believed that the development of a full vision system will be possible by extending the method. There are various reasons for this: firstly, the line drawing format is very restrictive - it is a low-level symbolic language for describing an image, but it is not a rich one. Secondly, it

has been found extremely difficult to write programs that can compute from an image, the near-perfect line-drawings that scene-analysis programs require. Thirdly, the knowledge contained in, for example Waltz's program, is in a certain sense too compiled to be used flexibly. The structure of the program makes it very efficient as a module, but rather unsuitable for interacting with other kinds of information - about the likely sizes, shapes, and positions of the objects in view. There is, for example, no way in which a description of the form "small object in the left foreground" can be used to help the analysis of the scene: it can be used only after the algorithm has run successfully, in which case its help was unnecessary. This is not an argument against the principle of modular design: the computational and evolutionary advantages of modularity need no further emphasis here, and one needs only to point to the work of Julesz (1971) to find an example of a nearly independent and impenetrable computational module in the human visual system. It simply means that the particular modularity represented by the scene-analysis approach constitutes an inappropriate subdivision of the total task; and I doubt whether this could have been predicted beforehand with confidence.

The more complex issues of visual analysis must be reserved until some earlier ones have been dispatched, so let us turn our attention to the development of programs that are capable of producing a line-drawing from an intensity array. The experience of Horn (1973) and of Shirai (1973) is that this process is extremely difficult to automate reliably. Unless the surfaces in the scene are specially treated, and the lighting

is arranged with some care, free-standing line-finders tend to fail. This led to the belief that the idea of a free-standing low-level vision module may in fact be unrealistic, and Shirai's program includes considerable scope for the guidance by higher-level processes of the lower-level ones. I shall argue in the accompanying articles that an almost free-standing unit is in fact realisable: the two factors that make it so are firstly, that changes in intensity are not the only useful cues in an image; other information, like discontinuities in the intensity gradient often provide useful, and sometimes the only, cues that an object boundary is present. Line-finders that ignore such information necessarily fail.

The second factor is that in order to make full use of such nuances in the image, the vocabulary in which the image is expressed must be expanded to include terms that describe them. This is an example of what seems to be a general principle of recognition systems, and one might call it the "art of the weak hypothesis". It is frequently the case during recognition that there are a number of possibilities for the interpretation of a particular datum, but that there is not yet sufficient evidence to distinguish between them. Some situations require complex computations to decide between the options - the disparity computation is one such example; and there may be another class of examples where a rather specific interaction must be allowed between the data and the goals of the computation (see Marr, 1975a). But there also seem to be many occasions where the selection of one of several options will not be possible until several steps further on. In such cases, one

should only as a last resort become committed to one of the possibilities, because of the damage that knowledge associated with that possibility and not with the others can subsequently do. In a natural language parser, one can simply wait and see (Marcus 1974); but in vision, the sheer volume of information effectively rules this out because of memory limitations. The description must be produced now, and it must be correct.

A fast straight-through recognition process will therefore have be based on conservative principles. Nothing can be assumed unless it is reasonably certain, and adequate descriptions of the data have to be available at each level using symbols at that level. If a system is unable to describe the data at any stage without using concepts that imply more than can reasonably be asserted at the time, then the system does not have enough concepts for fluent recognition: and the way to achieve fluency is to increase its vocabulary until it does.

One other factor has contributed significantly to the distraught attitude of computer vision projects to their low-level problems. It is that the analysis of a reasonably sized image in real time requires prodigious computing power, probably four orders of magnitude greater than that of a conventional general-purpose machine. This forces computer vision research into one of two paths: one can either accept the limitations of present machines, and expend one's energy and ingenuity on devising fast, specialised routines without any expectation that they will generalise. All of the current advanced automation projects take this path, and in many cases the prevailing conditions can probably be

controlled adequately: the great success of insect visual systems is but one piece of evidence for the value of special-purpose mechanisms. A second approach is to look for a more general-purpose scheme, in full knowledge that it will run slowly without special-purpose hardware, and that a working visual system will require several special mechanisms in addition to it. It is to the computational issues that arise in this second approach that this investigation is directed.

Finally, it is perhaps worth mentioning the considerable body of literature that comes under the general heading of picture-processing (see e.g. Rosenfeld (1969), and Rosenfeld (1973) for a bibliography). Some of this literature describes special purpose machinery designed to extract well-defined properties of an image; and some, called non-purposive vision, is more concerned with the design of a general purpose visual pre-processor. With those special-purpose mechanisms that work, no-one can quarrel: but the difficulty with much of the non-purposive vision literature is that unless a technique is based on some kind of theory, or forms a part of a larger computation, it is hard to evaluate it.

Some simple fallacies

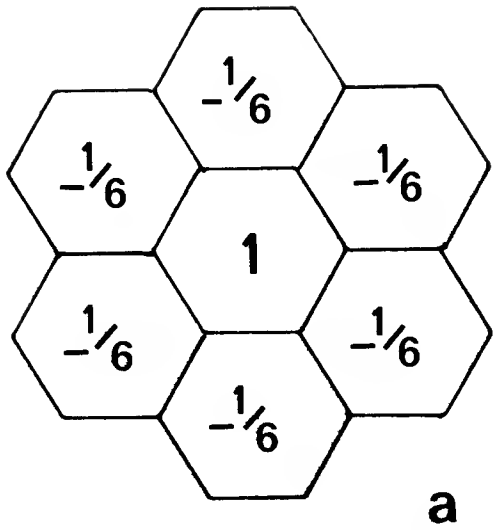
Let us return to the problem of computing symbolic assertions about an image from measurements made upon it. The application of a bar-shaped mask to an image does not, as we have seen, lead directly to an assertion about the presence of a bar in the image. The underlying point concerns the relation between computing the bar assertion, and the

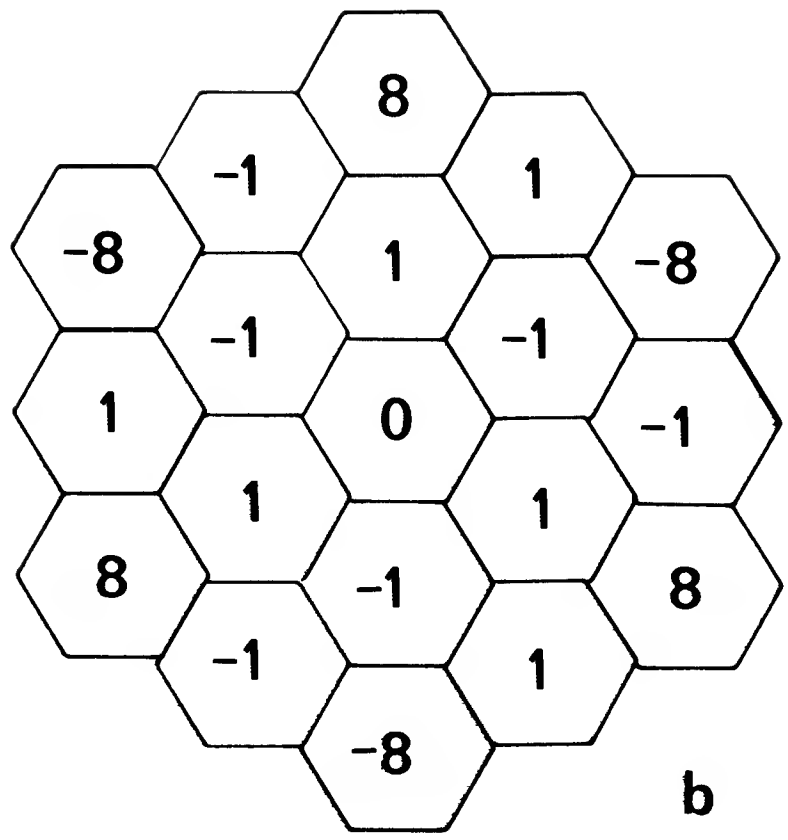
inverse transform of the original measurement, and it is a point of some importance. To illustrate it in another context, let us briefly consider the computation of an assertion about the presence of a blob in the visual field. A way of computing this assertion that immediately springs to mind is to take a circular mask, that has a positively weighted central region, and a negatively weighted surround, as illustrated in figure 3a. One might conjecture that a blob exists in the image at a point P provided that the mask gives a value k (say) at P, and a value -k/6 at the neighbouring points. These conditions are certainly necessary for the presence of a blob, but they are far from sufficient. Figures 3b, c & d give counter-examples of various kinds. In all of these, any additional intensity applied to the central point would give rise to the specified conditions at that point. The reason for the failure is that the inverse transform to that produced by a centre-surround receptive field depends critically on the boundary conditions. (It is the same one that was used by Horn (1974) and by Marr (1974d)). Any method that computes a blob assertion infallibly from the centre-surround measurements is in a sense computing part of this inverse, and so must take account of the boundary conditions. There will often be short-cuts: the frog's retina seems to use the conjunction of the condition defined above with the condition that the item in question be moving (Barlow 1953, Lettvin et al. 1959). This is apparently good enough, but it is not infallible, and should be easy to fool. Similar observations hold about the use of "tongue"-shaped and "corner"-shaped masks for the detection of tongues and corners. One might have imagined

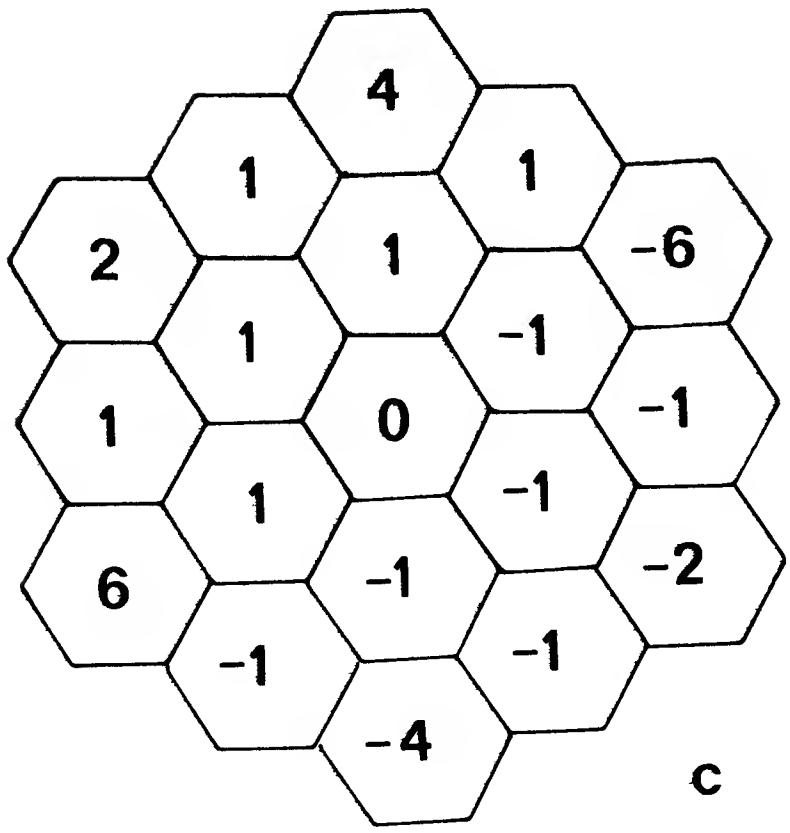
that a tongue-shaped mask could provide a sensitive detection mechanism for a tongue assertion, but if one tries it out, one quickly discovers that it is not. A tongue-shaped mask measures only the total intensity distributions over the positive and negative regions that it contains, and is insensitive to the distribution of intensity over these two regions. To include conditions on the distribution of intensity, one must take into account the structure of the inverse transform, and this depends in an intricate way on the boundary conditions. Using it to detect tongues in the image is therefore expensive. The overall conclusion that one may draw from these considerations is that for a general purpose vision system, the measurements from which assertions about the image are computed should be such that their inverse is easy to compute, (even though it is not explicitly obtained). In particular, the boundary conditions should be as unimportant as possible. The use of measurements that have an orientation sensitivity makes very good sense, because the boundary conditions for their inverses are 0-dimensional, (consisting of two points), rather than 1-dimensional as they are for a centre-surround organisation.

The importance and unimportance of linearity

Whether or not a low-level vision mechanism is linear or non-linear is a question that is of some practical, but less theoretical importance. The reason for its practical importance is that in higher mammalian visual systems, the initial convolution measurements are taken by a discrete grid of simple cells, scattered over the space of possible







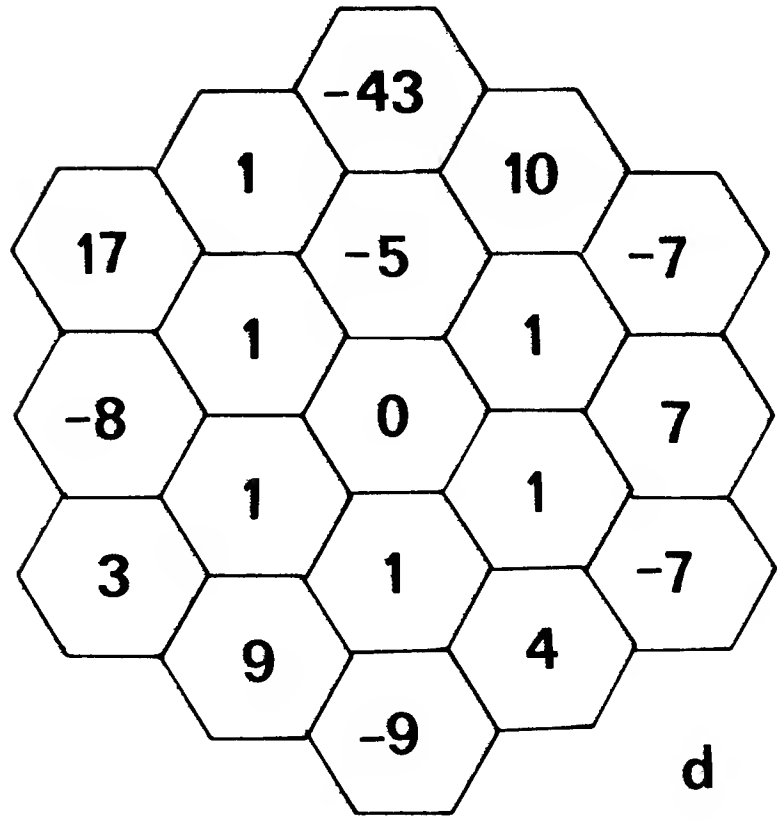


Figure 3

Figure 3. 3a illustrates the centre-surround mask that is described in the text. 3b, c, and d show intensity distributions that give a null response over the central seven positions. Hence any extra intensity placed at the central element would cause the simple algorithm for detecting a BLOB to assert that one was present there. These examples illustrate why such an algroithm would be inadequate.

positions, orientations, velocities, and so on. Because of this, values that lie between those of two nearby simple cells will be coded to some extent by interpolation. The same may be true of the representation of symbolic assertions, and of modifiers to those assertions. The advantages of linearity to this kind of interpolation are considerable. For example, the computation of fuzziness, a modifier associated with an edge (Marr 1975b), involves comparison of values derived from convolutions of the image with masks of different sizes. In a discrete hardware implementation of the process, the comparison must operate accurately for all nearby positions of the edge in the image. This means that the discrete units through which the values are represented should have a relation to one another that is fixed, and independent of the precise position of the edge in the image. For this to be possible, the system must be locally linear. The system cannot of course be globally linear, because the switch from one description to another is not a linear process and because few of the constraints imposed by the real world on the structure of the description are linear: but the argument does show that one may expect many of the components of the low-level symbolic descriptor to consist of locally linear sub-systems, between which switching takes place. One example of this is the observation that if a sine-wave grating is presented to one eye, and another of three times the frequency is presented to the other, they are perceived as nearly a square-wave if the phase relation is correct, and one can adjust the relative amplitudes while maintaining this perception: but if the phase relations are wrong, there is rivalry between the images (Maffei &

Fiorentini 1972). The system is locally linear, under the right conditions: but only locally.

Bindings

The low-level system that is described in the accompanying articles is based on descriptions of intensity changes, but there are clearly a number of other qualities that need to be associated with contours and regions in an image during its analysis. Obvious examples are velocity, orientation, position, binocular disparity, colour, descriptors of surface texture, and so forth. It is natural to suppose that these modifiers exist in the form of symbols that are bound to the appropriate shape descriptor. Thus the perception of the colour of a surface would correspond to the binding of a symbol for that colour to the symbol for the shape, rather than to a crude contour approximation analagous to that which occurs when one of the images in a colour-printing process is displaced relative to the others. (The implementation of the binding may of course use rough contour description as an addressing technique - see Marr 1974c).

The representation of such parameters as colour, disparity, and so on, by means of symbolic assertions that modify the underlying shape descriptors, makes a number of illusions easy to understand at least in general terms. The well-known waterfall effect, in which there is apparent movement without change in position, would correspond to an assertion of non-zero velocity becoming bound to the relevant item, but with the position binding (perhaps represented by using a naming scheme

for the directions round the viewer) remaining constant. The illusion of Fuchs (1923 figure 6) is another interesting example: nine small circular blobs are arranged in a 3 by 3 grid. The central blob is coloured yellow-green; the four corner ones are blue-green, and the four at the centres of the sides are yellow. If the central blob is seen as part of an x that includes the four corner blobs, it appears blue-green: if it forms part of a + with the blobs in the sides, it appears yellow. It seems to depend to what shape descriptor the colour value is bound. On close examination, the colour of the central blob can of course be diagnosed accurately.

The process of binding a value can be very simple, as it is in the case of binding a velocity value to a contour. At higher levels, a value may have to satisfy many prerequisites before it can be allowed to occupy a particular slot: but there are reasons for believing that extreme complexities may exist at the lower levels too. For example, the conditions under which a disparity value may be bound are extremely complex. I show elsewhere (Marr 1974c) how the constraints that the 3-dimensional world places on an acceptable assignment of disparity values to an image can be accurately represented as pre-conditions on the binding process. The same constraints cannot be accurately represented outside a discrete symbolic environment, which probably explains why methods that use grey-level correlation techniques for measuring disparity fail to work very well (Mori et al. 1973).

Ten computational problems in low-level vision

It is an interesting reflection on the primitive state of vision research that so very few of even the most basic low-level computational problems have been settled. The following list illustrates some of the more important ones, but it is not exhaustive, and some of the terms reflect a modularity that may turn out (as scene analysis did) to be inappropriate.

(1) A vocabulary for the primary parameters associated with an item in a (two-dimensional) image.

The primary parameters are velocity of movement, local orientation, and position. The computation of these parameters is the easier part: what is more difficult is the design of a vocabulary that allows efficient manipulation of the relation between these variables. For example, if position in the image is represented by a family of names for directions, and orientations are represented by another set of names for orientations, then the orientation of the line between two position names must be readily available. The eventual computation of three-dimensional predicates like ABOVE, BESIDE and so forth will also rely to some extent on the ease to which questions about two-dimensional configurations can be answered. The design of a low-level vocabulary for these primary parameters is therefore not as easy as it at first sight appears.

(2) Intensity changes

The low-level description of intensity changes in an image is

discussed at length by Marr (1974a and 1974b).

(3) Disparity

The work of Barlow et al. (1967), and of Julesz (1971), has provoked much interest in the characteristics of the human mechanism for assigning disparity values to an image (disparity being one kind of information that can lead to assertions about distance from the viewer). But there has been little progress in the study of parallel algorithms for performing the computation (see Marr 1974c).

(4) Lightness and colour

One of the more easily formulable problems is how one computes perceived colour. The method of Land & McCann (1971) (see also Horn (1974) and Marr (1974d)) appears to work acceptably in the model world of Mondrians for which it was developed: but because of the great extra complexities that arise in natural images there are some grounds for believing that this method needs considerable modification before it can be made generally reliable.

(5) Fluorescence and brilliance

The detection of sources of light in the visual field is an important ability, and the illusion of Evans (1974) seems to imply that the nervous system has an interesting and special method for doing it. Evans created two slides, transparent except for a small black square that appeared on the left of one slide, and on the right of the other. The positions of these squares did not overlap. The two slides were projected, one using red and the other using green light, and the two images were super-imposed. Most of the resulting image was yellow,

except for the two squares, one of which was red and the other, green. These squares appear to be more brilliant than the background.

Ullman (1975) has noted that measurements of local contrast gradient over a surface may be used to detect fluorescence. If a surface is illuminated by a source that is not too distant, the illumination gradient produced on the surface will be measurable. Two adjacent surfaces of constant but different reflectances will have the same local contrast gradient across them. If one of the surfaces is also a source, the local gradient there will however be smaller than on parts of the surface that are not sources, because a contrast measure involves dividing by the DC luminance in a region. This fact allows one to construct a sensitive and autonomous detector of fluorescence.

(6) Surface texture

One of the more vexed questions of low-level analysis has been the description of surface texture. Some kinds are not too difficult: glossiness is probably one of the simpler ones to detect, (and should probably be thought of like colour as a quality that is bound to the description of a surface). Textures like a grass lawn, or a woven fabric, seem at first sight to be inaccessible; but there are some grounds for believing that the analysis of textures with some order to them is not in fact too difficult (Marr 1975b). Very irregular patterns like cork-board seem to pose a rather different kind of problem; but the difficulty we have in describing them may simply reflect the fact that we do not possess very good internal descriptors for them ourselves, and so vacillate between several ways of describing it, none of which is very

successful. (Brodatz 1972 contains an interesting selection of visual textures).

Highlights, specularities, and their associated intensity peaks are difficult to study with conventional cameras because of overload problems. They are however of clear importance for two reasons. Firstly, the computation of qualities like metal, glossiness, glister, wetness, and oiliness seems to depend in part upon them; and secondly, if they can be recognised at a low level, their description can be excised from that of the rest of the image quite early on. This makes the recognition problem a great deal easier.

(7) The inference of boundaries

Changes in intensity or in intensity gradient are not the only useful cues that indicate the presence of an object boundary in an image. It is well-known that certain kinds of movement (Julesz & Hesse 1970), disparity, texture, "hatching" and various other changes can provoke the impression of a boundary in an image. For example, we have no trouble describing the overall shape of a tree in winter, when the only cues are bare and sparsely scattered twigs at random orientations. As far as I am aware, there has been no progress in the study of methods capable of achieving this.

(8) Transparency

One of the few areas that has been studied with some success in the question of diagnosing transparency (see Metelli 1974). There are however many awkward transparent objects - like a Coca-Cola bottle - which raise a number of problems simultaneously in a rather complex way:

Metelli is perfectly aware of the difficulties, and it is apparently not known how well his method performs in such cases.

(9) Symmetry

Julesz (1971 pp128-136) emphasised the fact that our perceptual apparatus seems to be sensitive to symmetries. Presumably, we do not contain complete representations of the important symmetric groups, but achieve our sensitivity by the use of a small number of tricks. The examples that he gives, though composed of random dot patterns, do not prove that we carry out a general symmetry test. It would be enough for those examples if we used the few, unusually shaped clusters of dots that appear by chance at various places in the image; this observation is however far from a precise definition of a working method.

(10) Figure-ground

The importance of figure-ground lies in its ability to select out of an image sub-parts that may usefully be included in a single descriptive unit. One would expect it to be implemented very soon after the computation of a very low-level symbolic description of an image. Given a low-level symbol asserting the presence of a vertically oriented EDGE with positive sign, there are two possible symbolic descriptions of this edge at the next stage: one can either regard it as a black border with the background on the right, or as a white border facing to the left. Selection of one of these is forced if it is known (e.g. from disparity information) that one side is closer than the other. Because figure-ground distinctions may be made in artificial nonsense images, it is generally thought that the underlying computation, though capable of

taking a number of aspects of the image into account, is a low-level one; yet there has been no attempt at a precise study of methods for carrying it out.

Some non-problems

Some of the traditional issues cease to be important problems for a system that relies from the start upon symbolic manipulations (see Minsky & Papert 1972). The translational invariance of a description is one such problem: the symbol LINE does not have a dependence on position and no problems arise in deciding whether two lines are the same kind of thing when they appear at different places in the image. The same holds true for LINEs at two orientations, but more complex issues become involved in the description of configurations of lines at different orientations. Provided that the symbolic representation of a configuration is constructed in terms of intrinsic predicates (e.g. PERPENDICULAR LINE1 LINE2), or relative to axes set up locally in the image, the representation will display rotation invariance: and one would expect many aspects of the description of three-dimensional shapes to be computed in such terms. The use of predicates that rely on the viewer's frame of reference (ABOVE, TO-THE-SIDE-OF) or on templates that consist of fixed point-configurations in the visual field, would however result in a representation with little invariance for large rotations.

A symbolic system will further not have size-constancy, in the traditional sense of using a size-invariant representation. The ability to recognise objects at different sizes arises for a variety of reasons.

Firstly, the grosser aspects of an object's description do not change very fast as the size of the image changes; and secondly, any competent recognition system must have available a large amount of information about the possible appearances of an object. The changes induced by lighting an object differently can be spectacular, compared with the changes in the description of its image produced by halving its size.

The scope of this article has purposely been limited to very low-level issues in a computational approach to vision, because its immediate purpose is to provide a background for the accompanying detailed investigations. Higher level issues are of course abundant, and in a sense more interesting than the somewhat pedestrian details with which low-level studies must concern themselves: but if the computational approach to vision has any value, it lies in its ability to separate out of the family of potentially useful ideas for visual processing those methods that actually work on real images. Until one tries to deal with a natural image, there is a tendency not to appreciate how complex and awkward they are. High level problems cannot be investigated experimentally without a fairly secure low-level system, because they need to be tested on natural data. A robust, low-level system is therefore a prerequisite for higher-level investigations.

References

Barlow, H.B. (1953) Summation and inhibition in the frog's retina. J. Physiol. (Lond.), 119, 56-68.

Barlow, H.B., Blakemore, C. & Pettigrew, J.D. (1967). The neural mechanism of binocular depth discrimination. J. Physiol. (Lond.), 193, 327-342.

Brodatz, P. (1966). Textures: a photographic album for artists and designers. New York: Dover Publications.

Evans, R. (1974). The perception of color. New York: John Wiley & Sons.

Fuchs, W. (1923). Experimentelle Untersuchungen uber die Aderung von Farben unter dem Einfluss von Gestalten ("Angleichungserscheinungen"). Zeitschrift fur Psychologie 92, 249-325.

Horn, B.K.P. (1973). The Binford-Horn LINEFINDER. A. I. Lab. Memo. 285.

Horn, B.K.P. (1974). On lightness. A. I. Lab. Memo 295.

Hubel, D.H. & Wiesel, T.N. (1962). Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. J. Physiol. (Lond.), 160, 106-154.

Jardine, N. & Sibson, R. (1971). Mathematical taxonomy. London: John Wiley & Sons Ltd.

Julesz, B. (1971). Foundations of Cyclopean Perception. Chicago: The University of Chicago Press.

Julesz, B. & Hesse, R.I. (1970). Inability to perceive the direction of rotation movement of line segments. Nature, 225, 243-244.

Land, E.H. & McCann, J.J. (1971). Lightness and retinex theory. J. opt. Soc. Amer., 61, 1-11.

Lettvin, J.Y., Maturana, H.R., McCulloch, W.S. & Pitts, W.H. (1959). What the frog's eye tells the frog's brain. Proc. Inst. Radio Engrs., 47, 1940-1951.

Mackworth, A.K. (1973). Interpreting pictures of polyhedral scenes. Artificial Intelligence, 4, 121-138.

Maffei, L. & Fiorentini, A. (1972). Processes of synthesis in visual perception. Nature (Lond.), 240, 479-481.

Maffei, L. & Fiorentini, A. (1973). The visual cortex as a spatial frequency analyser. Vision Res., 13, 1255-1267.

Marcus, M.P. (1974). Wait-and-see strategies for parsing natural language. A. I. Lab. Working Paper 75.

Marr, D. (1974a). The low-level symbolic representation of intensity changes in an image. MIT Artificial Intelligence Laboratory Memo 325.

Marr, D. (1974b). The recognition of sharp, closely spaced edges. MIT Artificial Intelligence Laboratory Memo 326.

Marr, D. (1974c). A note on the computation of binocular disparity in a symbolic, low-level visual prdcessor. MIT Artificial Intelligence Memo 327.

Marr, D. (1974d). The computation of lightness by the primate retina. Vis. Res. (In the press).

Marr, D. (1975a) The art of the weak constraint: a symbolic analog of solving simultaneous equations. (In preparation).

Marr, D. (1975b) Configurations, regions, and simple texture vision. (In preparation).

Metelli, F. (1974). The perception of transparency. Scientific Amer., 230, (April issue), 91-98.

Minsky, M. & Papert, S. (1972). Artificial Intelligence Progress report. M.I.T. A.I. Lab. Memo 252.

Mori, K., Kidode, M. & Asada, H. (1973). An iterative prediction and correction method for automatic stereocomparation. Computer graphics and image processing, 2, 393-401.

Rosenfeld, A. (1969). Picture processing by computer. New York: Academic Press, 196 pages.

Rosenfeld, A. (1973). Progress in picture processing: 1969-71. ACM Computing Surveys, 5, 81-104.

Shirai, Y. (1973). A context-sensitive line finder for recognition of polyhedra. Artificial Intelligence, 4, 95-120.

Ullman, S. (1975). M.I.T.Master's Thesis. (In preparation).

Waltz, D. (1972). Generating semantic descriptions from drawings of scenes with shadows. A. I. Lab. Technical Report 271.